Project Report

# GitHub URL

<https://github.com/p-renata/UCDPA_renata_pagac>

# Abstract

This Data Analytics assignment is focused on the best books listed by Goodreads. The reason for choosing this topic for my final assessment is because I love books and reading. I delved into the Goodreads dataset to explore the books that are considered the best by readers and critics alike. I explored the books by authors, date of publication, rating, number of ratings,and most used languages etc. Through this analysis, I aimed to gain a deeper understanding of the literary preferences of readers. This assignment is an opportunity to combine my passion for books and reading with newly acquired knowledge and skills in data analytics.

# Introduction

The objective of my analysis was to examine the correlation between books listed by Goodreads as the best books and those on Amazon as best-selling books.

The programming language used for this analysis was Python 3,, and the development environment was Jupyter Notebook (Anaconda3).

# Dataset

# I have chosen the following datasets for my analysis:

Goodreads books –

<https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>

Amazon –

<https://www.kaggle.com/datasets/phanee16/amazon-books-best-seller>

Implementation Process

1. Goodreads

* Importing required libraries
* Importing csv file using pd.read\_csv
* Cheching the first five rows of dataframe using iloc()
* Checking the firts two rows of data using .head(2)
* Checking for the numbers of rows and columns with .shape
* Checking the datatype of each column using .dtypes
* Removing the duplicates using .drop\_duplicates() (there is noo duplicates)
* Removing the rows where data is missing using .dropna()
* Checking what columns do I have in dataframe .columns
* Removing unwanted columns with .drop[]
* Changing the names of columns .rename()
* Seting column „ID“ as index .set\_index
* Saving dataframe into new csv

1. Amazon

* Importing required libraries
* Importing csv( retrived fro SQLite) file using pd.read\_csv
* Checking for the numbers of rows and columns with .shape
* Looking into the first 5 rows of data using .head(5)
* Checking the datatype of each column using .dtypes
* Removing the duplicates using .drop\_duplicates() (there is noo duplicates)
* Removing the rows where data is missing using .dropna()
* Checking for the sum of missing data in each column using .isnull().sum()
* Checking what columns do I have in dataframe .columns
* Changing the names of columns .rename()
* Saving dataframe into new csv

1. Analysis

* Importing required libraries
* Importing data from csv I saved earlier
* Droping unwanted columns
* Changing date form in „Publication\_Date“ column
* Sorting data by the date of publication from the highest year to lowest
* Checking for missing date in each column (after I noticed there are some) using .isnull().sum()
* In the column „Publication\_Date“ i had two missing data what I showed on heat map before a remove the columns with them
* Changing all the values to the integer in „Publication\_Date“ column
* Sorting the books by the publish year using boolean
* Using matplotlib to show the ratio of the books published in specific years
* Checking the column „Language“ to se what languages are apearing and how many times using .value\_counts()
* Replacin some of the values in „Language“ column using .replace()
* Using the matplotlib show the share of all other languages vs english language
* Applying groupby() function to group the data on language value
* Printing the first entries in all the groups formed using groupby.first()
* Finding the values contained in the specific group using groupby.get\_group()
* Showing the relationship between average rating and ratin counts using seaborn.histplot
* Finding out the first 20 authors with most occurences in the list on ste seaborn.barplot using value\_counts()
* Finding out which 10 books are most reted using seaborn.barplot
* Importing another dataset from csv (earlier saved from Amazon data)
* Merging Amazon dataframe with Goodreads on „Titles“ using inner join to keep the rows that apeas in both dataframes.
* As I had two columns for „Average\_Rating“ and „Rating\_Counts“ a made the new ones using mean() fuction to find average value betwen two columns
* Checking for all the names of columns .columns
* Dropping unwanted columns .drop()
* Change the name of column .rename()
* Making the new list that will be converted into a new column
* Checking for the columns after the change
* Accesing to specific column using df[column\_name]
* Accesing to first 5 rows of specific columns using df[0:5][[„Column1“, „Column2“]]
* Converting column „Price“ to numpy array
* Making new column for the prices increased by 5 percent
* Sorting data by price from the highest to the lowest
* Adding new column using if statment and boolean
* Looping through column
* Checking for the books published after 2010 using boolean
* Getting the value count of each genre variable in dataframe .value\_counts()
* Showing the ratio of genres using seaborn countplot
* Checking the average rating using seaborn.histplot
* Sorting all the books by numbers of rating and showing them using seaborn.barolot

# RESULTS FROM GOODREADS

# Chart 1: Missing data showed on heat map

# 

# Chart 2 : Showing number of the books published in different periods

# 

Chart 4: Comparing the share of English language with other languages

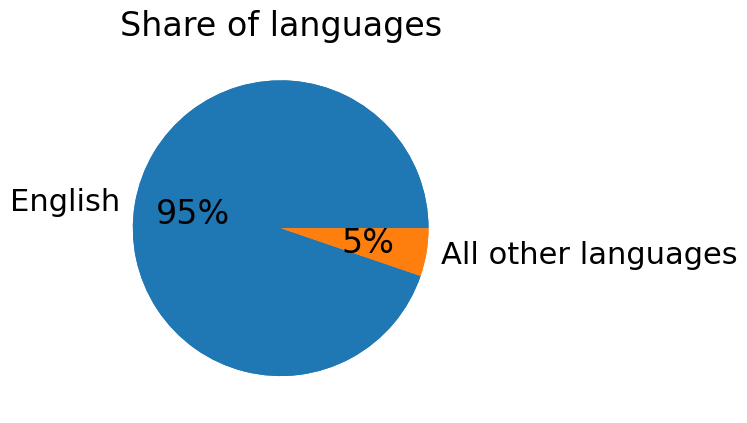


Chart 5: Comparing the share of languages without english language

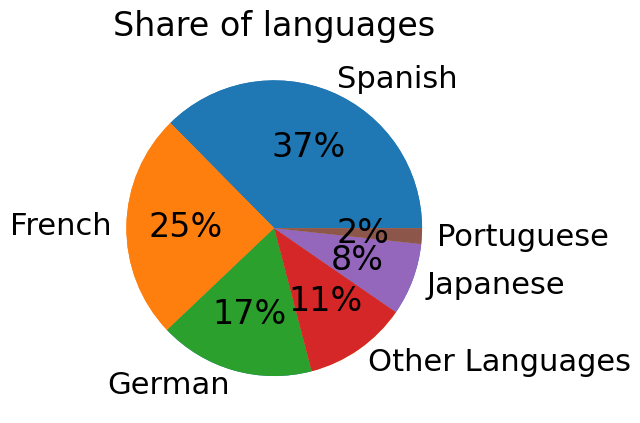


Chart 6: Relationship between average rating and rating counts

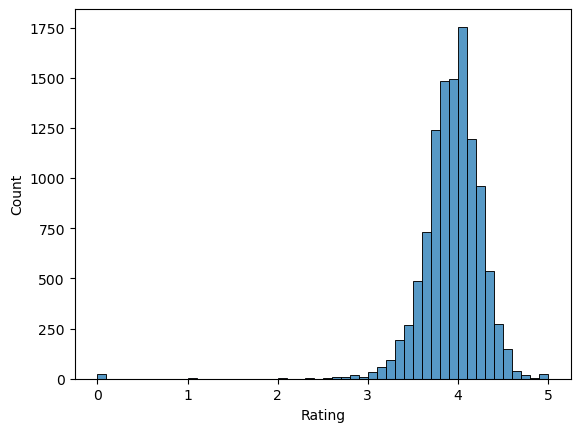


Chart 7: Most occuring authors listed by Goodreads

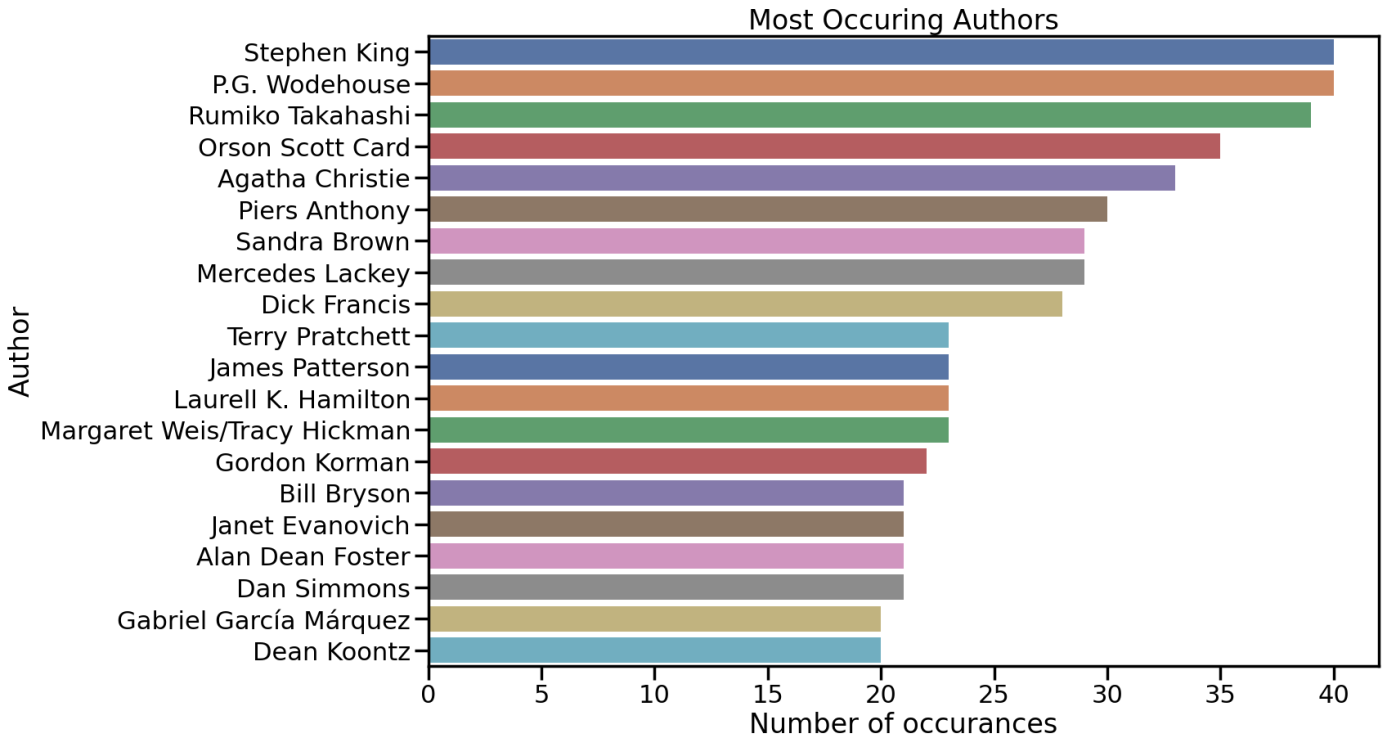
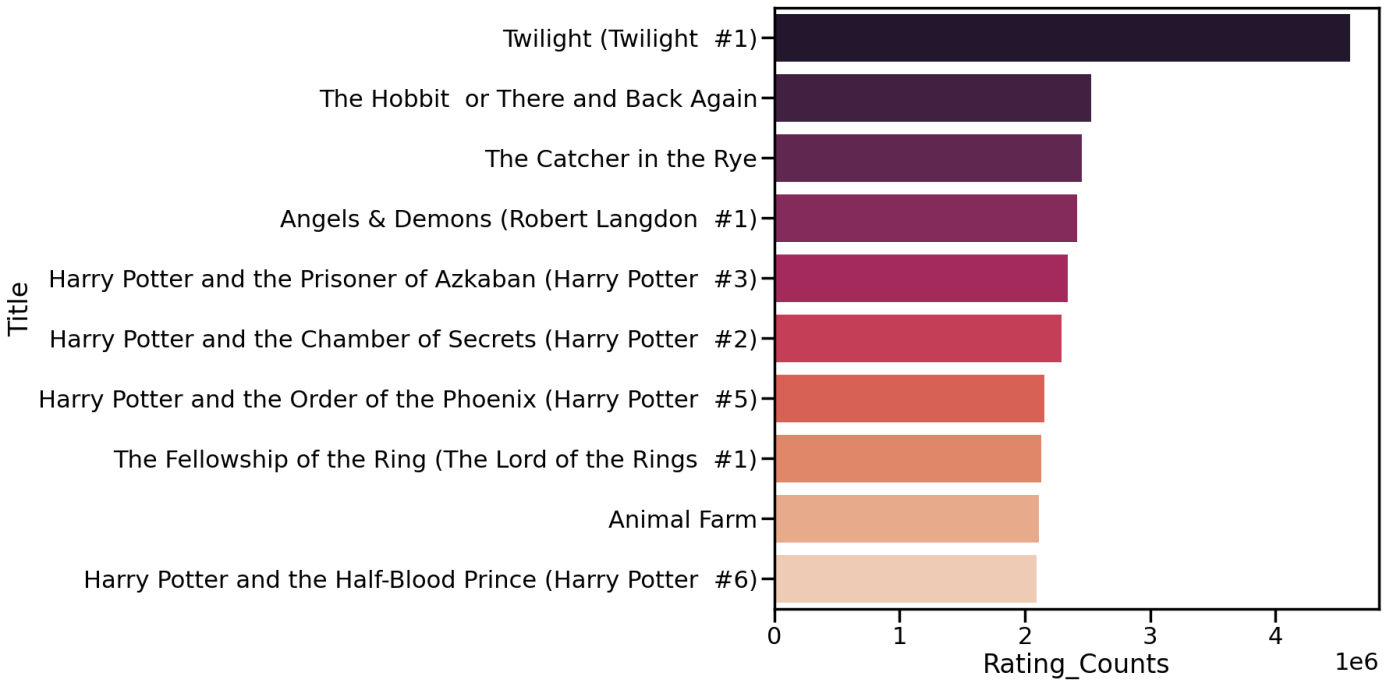


Chart 8: Most rated books listed on Goodreads



RESULTS AFTER MERGING GOODREADS AND AMAZON DATA

Chart 9: Showing the share of the books by genre

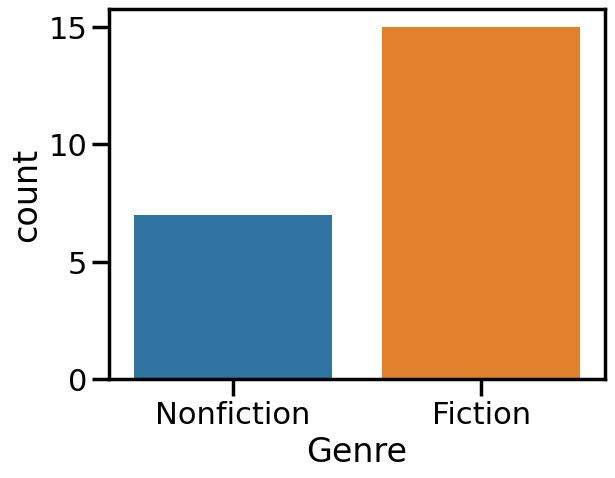


Chart 10: Relationship between average rating and rating counts

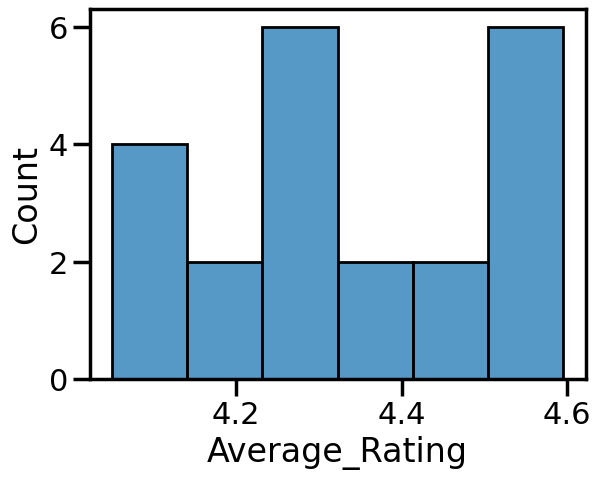
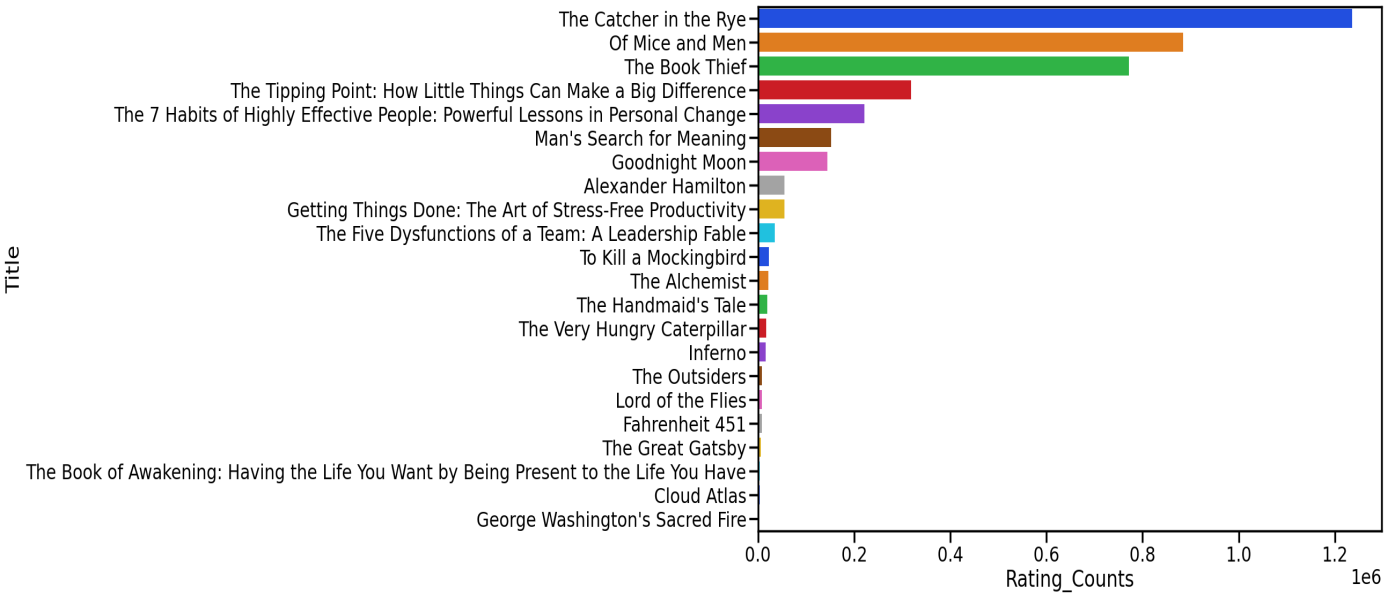
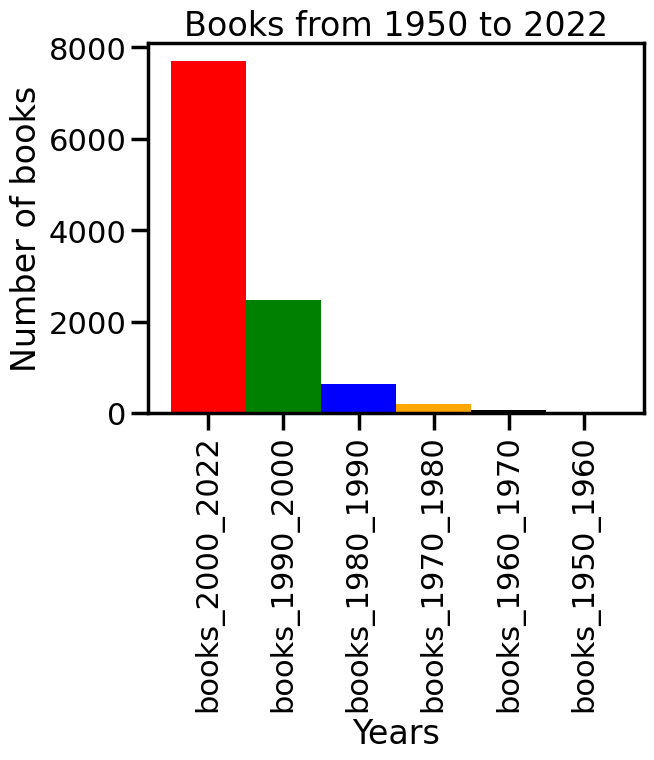


Chart 11: Most rated books found in both data (Goodreads and Amazon)



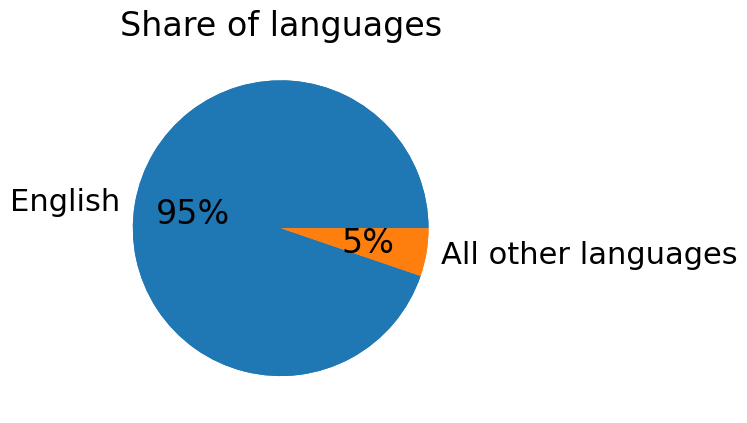
INSIGHTS

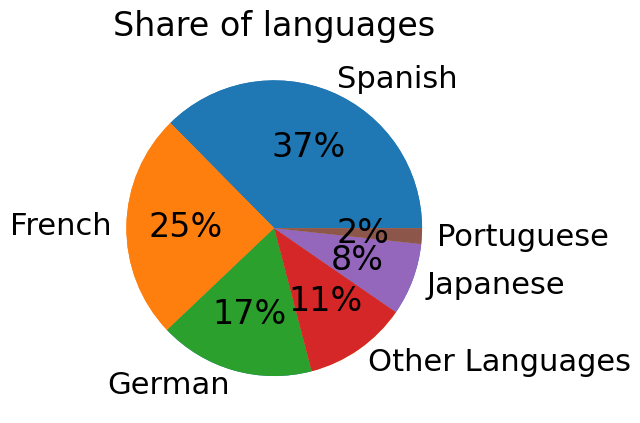
1: There is over 10,000 books listed by Goodreads, and it was interesting to find out that nearly 8,000 of them were published between 2000 and 2000. Over half of all the books were published since 1950. This data shows that in the recent years there is a significant increase in the number of books being piblished.

2: From the first chart we can see that English language is the most represented with 95 percent, while all other languages combined take up only 5 percent. This indicates that the majority of book listed by Goodreads are written in English.

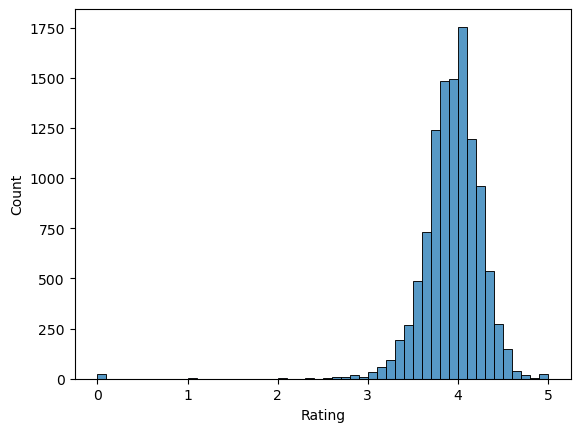
From the second chart, we can see which languages are the most represented after English. Those are Spanish, German, French, Portuguese, and others.

This data shows that although English is the dominant language in terms of represenntation on Goodreads, there are still a significant number of books listed in other languages as well.

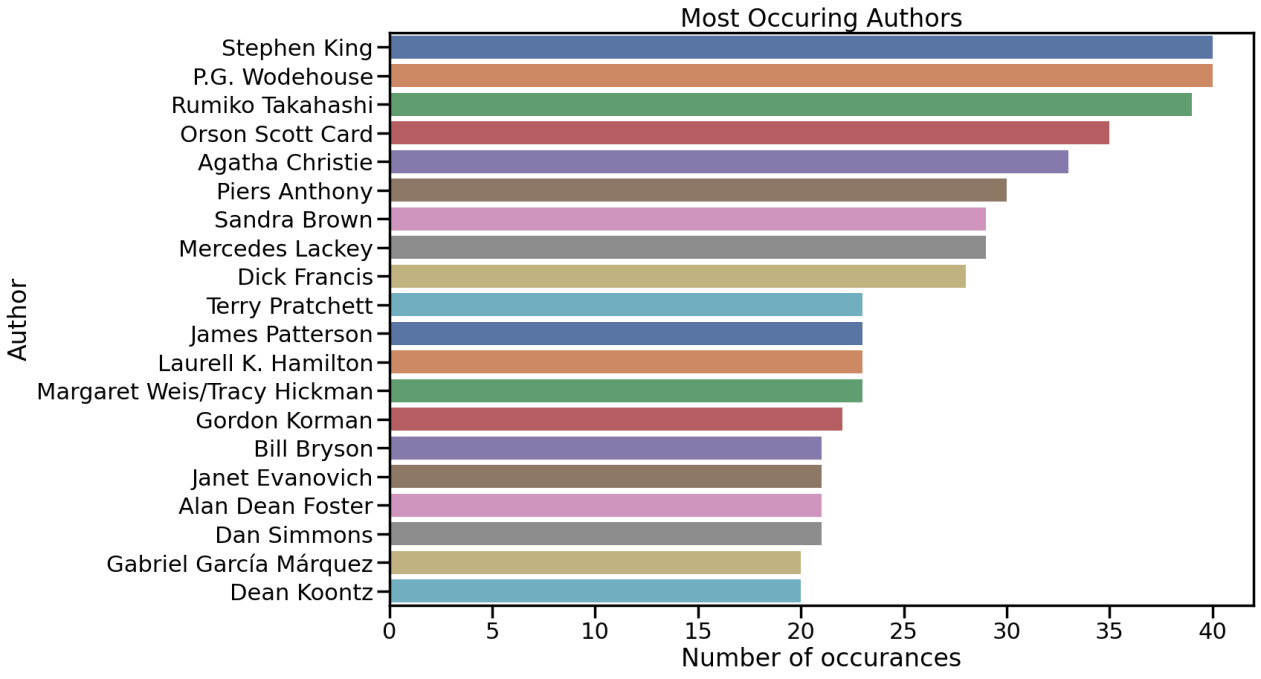




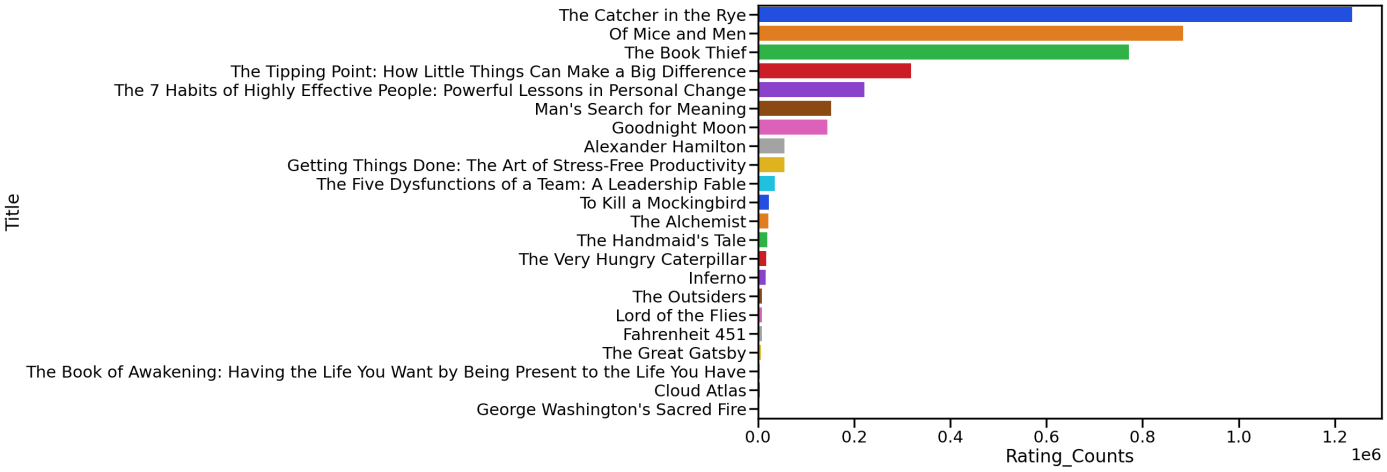
3:We can see that nearly all the books are rated between 3 and 5 . The average rating of the books on Goodreads is around 4 with over 1500 rating counts. Only a small percentages of books have rating 3 or less., which shows that the books listed on Goodreads are generally of high quality and well-reviewed. Additionally, it is visible that books with high rating counts have also high ratings, which suggest a correlation between popularity and quality of the books in this dataset.



4: We can notice that some authors are occurring over 20 times on the Goodreads list. This suggests that these authors have a significant number of books that are considered to be best books by the readers of Goodreads. Having one book on the list of best books is a big success, but having 20 books or more is truly amazing. This could indicate that these authors are highly respected in the literary community and have a consistent writing style, subject matter or genre that appeals to readers.



5. After going through over 10,000 books in the Goodreads dataset it was very surprising to find out that only a few books were listed on Amazon as best-selling books. As we can see in the chart, only 22 of them are listed there. This suggests that while many books are well-liked by readers and have a high rating on Goodreads, they may not necessarily be popular or well-selling on other platforms such as Amazon.



**Machine Learning potential on these Datasets**

Further analysis could de carried on the book dataset in the area of Machine Learning.

Regression is a type of analysis that aims to predict a continuous variable based on one or more input variables.

Examples of regression tasks that can be performed on a book dataset:

* Predict the sales of the books based on factors such as genre, language or author
* Predict the rating of a book based on reviews or authors
* Predict the price of the book based on factors as genre or author
* Predict popularity or users rating

Classification is a type of machine learning that aims to predict a categorical variable based on one or more input variables.

* to classify books into different categories (as genres)
* to classify books based on their authors

# References

<https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>

https://www.kaggle.com/datasets/phanee16/amazon-books-best-seller